Beszélő fej

Czap László[1], Mátyás János[2]

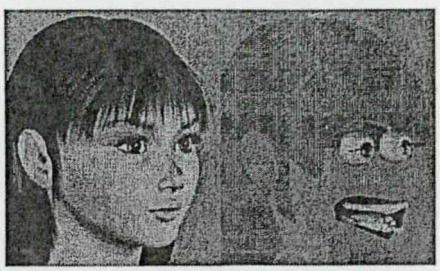
[1] Miskolci Egyetem, Villamosmérnöki Intézet, Automatizálási Tanszék
3515 Miskolc, Egyetemváros
czap@mazsola.iit.uni-miskolc.hu
[2] Észak-Magyarországi Regionális Munkaerőfejlesztési és Átképző Központ
3518 Miskolc, Erenyő u. 1.
matyasj@mail.erak.hu

Abstract. Magyar nyelvű, vizuális szövegfelolvasó fejlesztéséről számolunk be cikkünkben. Az animáció háromdimenziós fejmodell mozgatásán alapul. Az artikuláció kialakításához felhasználtuk a fellelhető hangalbumok anyagát, a dinamikus vizsgálatnál saját vizuális beszédfelismerési kutatási eredményekre támaszkodtunk. A koartikulációs hatások figyelembe vételéhez a jellemzőket domináns, rugalmas és határozatlan osztályokba soroltuk, ezek alapján határoztuk meg a mozgásfázisok közötti interpolációt. A természetesség javítása érdekében többek között álvéletlen fejmozgásokat és pislogást programozunk. A szemöldök mozgatása fontos szerepet játszik a gesztus kialakításában. A fejmodell működtetése során megvalósítjuk alapérzelmek kifejezését is. A cikk végén kijelöljük a továbbfejlesztés irányait.

1 Bevezetés

Mindenki előtt ismert, hogy a beszéd érthetőségét javítja, ha látjuk a beszélő személy arcát, ezzel együtt az artikulációját. Ez a vizuális információ különösen sokat segít zajos környezetben és hallássérültek esetében. A gépi beszédkeltés jól kidolgozott rendszereinek természetes kiegészítője a mesterséges beszélő fej. Az arcanimáció megvalósítása a beszédartikuláció modellezésére mindössze két évtizeddel ezelőtt kezdődött. A mai szemmel kezdetleges eszközökkel végzett első próbálkozások a vizuális beszédszintézis kezdetét jelentették. A 3D modellezés fejlődése, a számítástechnikai eszközök kapacitásának robbanásszerű bővülése és a természetes artikuláció analízise életszerű, fotorealisztikus finomságú modellek kidolgozását tette lehetővé.

Az elmúlt évtizedben a terület dinamikusan fejlődött, egyre több alkalmazás jelenik meg. Az ember-gép kapcsolatban új távlatokat nyithat az audio-vizuális beszédszintézis és beszédfelismerés. Dialógus és oktató rendszerekben az érthetőséget és az attraktivitást nagyban javítja a beszédanimáció. Multimédia alkalmazásokban a virtuális bemondó vagy szereplő tágítja a művészi szabadság határait. Hallássérültek beszélni tanítását segítheti a helyesen artikuláló virtuális bemondó, amely átlátszó arcával a természetes beszélőnél jobban megmutatja a hangképzés részleteit.



1. ábra Fotorealisztikus és transzparens megjelenítés

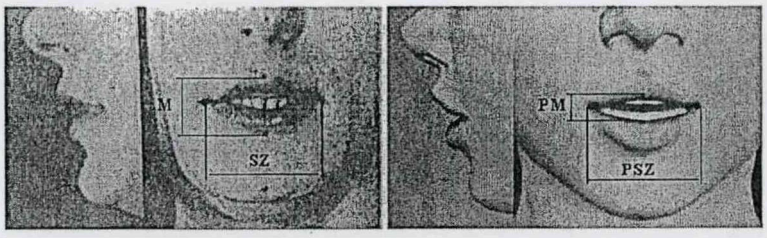
Hangvezérelt beszélő fejek fejlesztésén dolgoznak hallássérültek segítésére távközlési alkalmazásokban. A fejlett magyar nyelvű akusztikus beszédszintézis mellett hiánypótló célzattal kezdtünk vizuális beszédszintetizátor fejlesztéséhez.

2 A beszédanimáció

Az első működőképes vizuális beszédszintetizátorok kétdimenziós modell mozgásfázisainak előállítására épültek, kezdetben előre tárolt képek előhívásával. A kulcskeretek közötti fázisokat gyakran morfológiai módszerekkel állították elő. A kétdimenziós modell nem teszi lehetővé a természetes fejmozgások, a beszédet kísérő gesztusok és érzelmek kifejezését. A testmodellezés fejlődése a háromdimenziós modellezésre terelte a kutatók figyelmét. A 3D modellek egyik típusa az arcizmok megfeszítésével szimulálja az arckifejezéseket. Az ilyen modellek valósághű eredményt nyújtanak, de a kívánt arckifejezés előállítása rendkívül számításigényes és a valóságos izomtónusok nem mérhetők. Ma még ígéretesebb a pusztán felületi hatásokat utánzó, a bőrszövettel borított drótváz alakítására alapozott animáció. Ennek paraméterei megfigyeléssel, vagy képfeldolgozási módszerekkel természetes beszélők képeiről leolvashatók. [1] Minden modell mozgatásánál külön figyelmet kell fordítani a jellemzők összehangolt változtatására, mert könnyen természetellenes hatás alakulhat ki. Például az alsó fogsor és az áll független mozgása groteszk hatást kelt.

2.1 A beszéd vizuális alapegysége

A beszéd legkisebb akusztikus egységének, a fonémának vizuális megfelelője, a *vizéma*. A vizémák készlete szűkebb a fonémákénál, hiszen néhány fonéma artikulációja vizuálisan megegyezik. Nem látható pl. a zöngésség, de a képzés helyében megegyező, időtartamban vagy intenzitásban eltérő hangok is azonos artikulációs mozgásokkal jelennek meg. A hangképző szervek jellemző helyzete magyar beszédhangokra megtalálható alapvető munkákban [3], [4], [5]. A 2. ábrán példát mutatunk be arra, hogy mennyire hasonló egy hagyományos labiogram [4] és egy 3D-s beszélő fejen beállított ugyanazon hangra jellemző artikuláció.



2. ábra A minta fotolabiogram és a renderelt 3D fejmodell

A magyar beszédhangok vizéma készletét a [4]-ben megadott mintaszavak artikulációs jellemzőiből alakítottuk ki. Az eredményt az 1. táblázat mutatja, a hangokat a magyar helyesírási betűképükkel jelöljük.)

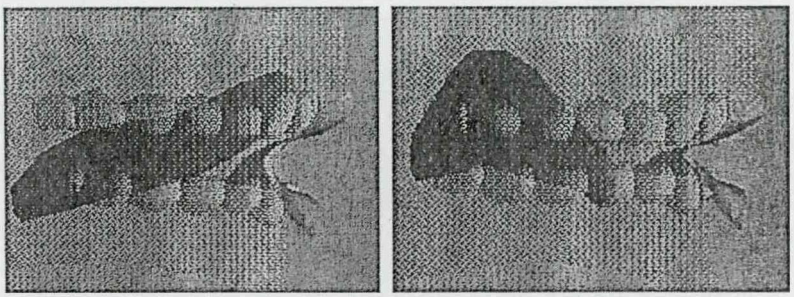
1. táblázat A magyar nyelv vizéma készlete

Magánhangzók	Mássalhangzók
e	b, p, m
é	f, v
I	t, d, n
ö, o	r
ü, u	sz, z, c, dz
á	l
a	s, zs, cs, dzs
	ty, gy, j, ny
	k, g
	h

Néhány megjegyzés a vizémák osztályozásához:

- a csoportosítás elsősorban ajakforma alapján történt, a nem látható nyelvállás eltérő lehet (pl.: o-ö, u-ü)
- a nem jelzett hosszú magánhangzók a rövid párjuknál szűkebb szájnyílással vannak jelen
- az artikuláció előállításához ennél bővebb készlettel dolgozunk

Az eddig megjelent beszédhangok atlasza [3], illetve magyar hangalbumok [4, 5] alapján meghatározhatók a vizémák legfontosabb paraméterei, ezekből alakul ki az a kulcskeret (keyframe) készlet, amely az artikuláció kiindulási alapja [6]. A legfontosabb jellemzők az ajkak és a nyelv működtetéséhez tartoznak. Az alapvető ajak jellemzők: nyitás (tág-szűk), szélesség (széles-keskeny), Az ajkak nyitása szoros összefüggésben van az állkapocs mozgásával (nyitott - zárt). A száj szélessége tehát az ajaknyitással és az ajakkerekítéssel, illetve az ajakréssel, áll összefüggésben. Az állkapocs helyzete a nyitás mellett a fogak láthatóságával is összefügg. A nyelvállást (2. ábra) a nyelv függőleges helyzete (fent-lent), vízszintes mozgása (elül-hátul), hajlítása (domború-homorú), és a nyelvhegy formája (széles-keskeny, vékony-vastag) befolyásolják.



3. ábra Jellemző nyelvállások: baloldalon az n-re, jobbra a k-g hangokra

A statikus jellemzők alapján beállíthatók a beszédhangok állandósult szakaszára jellemző artikulációs paraméterek, kulcskeretek.

2.2 Dinamikus működés

A folyamatos magyar beszéd dinamikus jellemzőinek átfogó leírása még várat magára. Az analízis során a hangalbumokban található pillanatképek korlátozottan használhatók, és csak a mintaszavakra vonatkoztathatók. A dinamikus analízis másik forrása a saját, vizuális beszédfelismerési kutatásaink során nyert eredményekből összeállított adatbázis [7]. Ebből származnak az ajkak nyitásának és szélességének időbeli változására vonatkozó adatok, valamint a nyelv és a fogak láthatóságát reprezentáló intenzitás faktor, a szájüregre vonatkozóan. Ezek a kulcskeretek közötti interpoláció megválasztásában nyújtanak segítséget.

A koartikulációs hatások figyelembe vételéhez túl kellett lépnünk az úgynevezett „keyframe" modellen. A vizémák minden jellemzőjét (például ajak- és nyelvállások) osztályoztuk dominans jellegük alapján. Egyes paraméterek a környezettől függetlenül felveszik jellegzetes értékeiket, mások a környezetükbe simulnak. A vizuális beszédfelismerés adatainak szórása alapján a vizémák jellemzőit három kategóriába soroltuk:

- *domináns* – nem enged koartikulációs hatásoknak
- *határozatlan* – a környezete alakítja ki az adott jellemzőt
- *rugalmas* – a környezete befolyásolja az adott jellemzőt

Példaképpen megadjuk a vizémák ajakformára és a nyelv vízszintes helyzetére vonatkozó csoportosítását (2. táblázat):

2. táblázat Dominancia jellemzők az ajakformára nézve

Domináns	magánhangzók, s, zs, cs, dzs
Határozatlan	t, d, n, r, l, ty, gy, j, ny, k, g, h
Vegyes	p, b, m, f, v, (szájnyílás domináns, szélesség határozatlan) sz, z, c, dz (szájnyílás rugalmas, szélesség határozatlan)

3. táblázat Dominancia jellemzők a nyelv vízszintes helyzetére nézve

Domináns	t, d, n, r, l, ty, gy, j, ny, s, zs, cs, dzs, sz, z, c, dz
Rugalmas	magánhangzók
Határozatlan	p, b, m, f, v, k, g, h

A dominancia beállításai a paraméterek interpolációs szintjét határozzák meg. A további módosítások – pl. hosszú magánhangzóknál állandósult szakasz beiktatása – finomítják az artikulációt.

3 A természetesség javítása

A beszélő természetes fejmozgását, mimikáját hírolvasó bemondók felvételein tanulmányoztuk. Ennek nyomán álvéletlen mozgásokat, például visszafogott bólogatást, a fej enyhe oldalra billentését és átlag körül szóródó pislogási periódust alkalmaztunk. A prozódia tükröződése a fejmozgásban, illetve az arcmimikában nehezen algoritmizálható, így pl. a mondathangsúly kifejezése nehézségekbe ütközik. Az intonáció azonban felhasználható a szemöldök mozgatásának vezérlésére. A mondathangsúlynál is emelhető a szemöldök. A szemmozgást a fejmozgás korrigálására használjuk, hogy a tekintet ugyanarra a pontra szegeződjön, egyéb szemmozgatás kézi beavatkozást igényel. Dialógus rendszerekben a szerepváltást segíthetik a gesztusok, az értő figyelést a szemöldök emelésével jelezhetjük, bólogatással is viszszaigazolhatjuk figyelmes hallgatásunkat. Ezek a műveletek egyelőre manuálisan állíthatók be.

3.1 Az érzelmek kifejezése

A beszéd multimodális jellegéhez hozzátartoznak a gesztusok is. A testbeszéddel árnyaljuk mondandónkat, megerősítjük vagy éppen cáfoljuk verbális üzenetünket. Arcanimációs rendszerünkben az arckifejezések érzelmi töltését próbáltuk meg algoritmizálni és programozni. Az Ekman [8] által meghatározott hét érzelem közül választhatunk: semleges, haragos, ellenszenves, szorongó, boldog, szomorú, meglepett. Erre láthatunk példát a 4. ábrán.

4. ábra Ellenszenves és boldog arckifejezés

4 Összefoglalás és kitekintés

A cikk többéves kutató-fejlesztő munka eredményét ismerteti. A cél vizuális szövegfelolvasó rendszer kialakítása. A fejlesztés jelen fázisában az artikuláció dinamikus jellemzőinek további finomítását végezzük. A természetes vagy gépi beszédhez a szinkronizálás még nem teljesen automatikus, a következő feladatunk ennek megoldása. Több fejmodell működtetéséhez egységes leíró nyelvet kívánunk kialakítani, hogy egy új modellhez csak a modellfüggő átalakításokra legyen szükség. Igy az esetleges fejlesztéseket is csak itt kellene megvalósítani, nem minden modell vezérléséhez külön-külön. A fejlesztőrendszerünk a beszélő fej videó anyagát hosszadalmas számításokkal állítja elő, ami több órás feldolgozási időt is jelenthet. Jelenleg csak olyan alkalmazásokra gondolhatunk, ahol előzetesen rögzített üzeneteket jelenítünk meg. A nagy számításigény miatt szövegfelolvasásra is alkalmas rendszerünk jelenleg csak kötött szótáras alkalmazásokban használható. Reményeink szerint a real-time animáció a közeli jövőben szuperszámítógépek nélkül is megvalósítható lesz és ezzel a tényleges virtuális bemondói, felolvasói alkalmazások is megvalósíthatók lesznek.

5 Köszönetnyilvánítás

A kutatást az Informatikai és Hírközlési Minisztérium az ITEM program keretében támogatta 345 regisztrációs szám alatt.

Irodalomjegyzék

1. Massaro, D.W.: Perceiving Talking Faces. The MIT Press Cambridge, Massachusetts London, England (1998) 359-390
2. Bernstein, L.E., Auer, E.T.: Word Recognition in Speechreading. Speechreading by Humans and Machines. Springer-Verlag, Berlin Heidelberg, Germany, 1996, 17-26
3. Molnár József: A magyar beszédhangok atlasza Tankönyvkiadó, Budapest, 1986
4. Bolla Kálmán: Magyar fonetikai atlasz. A szegmentális hangszerkezet elemei Nemzeti. Tankönyvkiadó., Budapest. 1995.
5. Bolla Kálmán: Magyar hangalbum : A magyar beszédhangok artikulációs és akusztikai sajátságai MTA Nyelvtudományi . Intézet., Budapest. 1980.
6. Mátyás János.: Vizuális beszédszintézis. Diplomaterv Miskolci Egyetem 2003.
7. Czap, L.: Lip Representation by Image Ellipse. ICSLP 2000 Bejging, China, Proceedings Vol. IV. 93-96
8. Ekman, P., Friesen, W.: Facial Action Coding System Consulting Psychologists Press. Inc., 1978.